External Independent Peer Review of Fisheries Stock Assessments for Arrowtooth Flounder, Flathead Sole and Kamchatka Flounder

By

Robin Cook

Prepared for
Center for Independent Experts
Independent System for Peer Review

Contents

Executive Summary

i. All three species assessments are based on a similar range of data types and modelling approaches. The principal data comprise biomass estimates obtained from the EBS shelf survey, the slope survey and the Aleutian Islands survey. All three surveys appear to be conducted to a high standard. CVs for the surveys are minimal estimates of the uncertainty and studies to obtain more comprehensive estimates would help in ensuring data in the assessments are weighted appropriately in the objective function.

ii. As the surveys take place on commercial vessels that change periodically, an investigation to vessel effects would help in understanding if changes in catchability are of material importance. The assessment approach is to assume survey catchability is constant, and in some cases known, which makes understanding the effect of vessel changes particularly important.

iii. There is much to be gained by being able to combine estimates from the EBS and slope surveys since this is a contiguous area and species are distributed in both. A study to find ways of inter-calibrating the survey biomass estimates might help in avoiding the need to make ad hoc assumptions in the assessment, especially where this relates to survey selectivity.

iv. An overview of the flatfish age reading program showed that good quality control procedures were in place to ensure consistent interpretation of growth rings though older fish are subject to higher aging errors. More otoliths are collected than are actually aged and increasing the number of specimens read is likely to be a worthwhile investment, especially for the arrowtooth and Kamchatka flounder fisheries.

v. The assessments are partitioned by sex potentially increasing the number of selectivity parameters to be estimated. I would recommend that efforts are made to reduce the number of parameters by consolidating selectivity curves were possible. Modelling fully selected F as a time series may also help to reduce the effective number of parameters further. The use of an information statistic such as AIC may help in model selection to avoid over-fitting the data.

vi. A number of choices are made by analysts which affect model fitting that include the imposition of penalty functions on some parameters and emphasis factors to weight the data. This tends to result in a lack of transparency in what the model is actually estimating and whether the data are weighted appropriately. Greater transparency could be achieved through the adoption of a more formal Bayesian framework where priors are explicitly stated and the likelihood is uncontaminated by *ad hoc* emphasis factors. Comparing prior and posterior distributions of the parameters would help understand the true information content in the data and identify where additional model constraints are justified.

vii. Natural mortality and survey catchability are problem quantities in the assessments as they are not well known and are difficult to estimate from the assessment data. There is no simple solution to this problem though it is probably unwise to either fix or estimate both simultaneously as they are often confounded. As the stock trajectories in these assessments do not show much contrast, it is likely to be better to estimate catchability and fix M. Alternatively where F is negligible, using survey age compositions could be used to get values of Z using the ratio of log numbers by cohort and this would approximate M.

viii. The arrowtooth and Kamchatka flounder assessments are closely linked because the species are caught in the same fishery. The species were only recorded separately in the catches from 2011. In view of these factors there may be advantages in performing a combined assessment where the species are treated as separate but subject to a similar fishing mortality and fit the model to combined data where the species are not discriminated.

ix. While recognizing that it potentially increases the amount of assessment work, it may be worth performing separate assessments for the Aleutian Islands. This would overcome some the problems of combining surveys that have different catchabilities and selectivities. In particular, it would avoid driving assessments with the EBS survey which covers the majority of the biomass, but which may not characterize size and age compositions in the AI area appropriately.

x. The flathead sole model was reproduced in Stock Synthesis. The strong decline in fishing mortality in the ADMB model largely disappears in the SS model. Estimates of fishing mortality prior to 1990 are highly sensitive to the choice of selectivity assumption. This is indicative that the early period is not well determined. Modelling selectivity parameters as a time series may help with model fit and avoid over-parameterization.

xi. The science reviewed represents the best information available within the limitations of the data. The modelling approach uses advanced statistical methods that are close to state of the art. The survey data are of high quality though the fishery data are limited in scale and lack age information for two of the species.

Background

1. The National Marine Fisheries Service (NMFS) is mandated by the Magnuson-Stevens Fishery Conservation and Management Act, Endangered Species Act, and Marine Mammal Protection Act to conserve, protect, and manage the nation's marine living resources based upon the best scientific information available. The Alaska Fisheries Science Center's (AFSC) Resource Ecology and Fisheries Management Division (REFM) requested an independent review of the integrated stock assessments that have been developed for three Bering Sea flatfish species; arrowtooth flounder, flathead sole and Kamchatka flounder. The assessments are performed using Automatic Differentiation (AD) Model software that uses survey abundance data, survey, fishery age, and length composition data with a harvest control rule to model the status and productivity of these stocks and set quotas.

2. An independent review of assessments of the flatfish was organised to assist in the development of final assessments that will go forward to inform the management process. The review considered preliminary assessments for each stock that comprised updates of models used in previous years and new models configured to address weaknesses or develop new approaches to the assessments. The review was intended to help guide further development of the assessment models and was not a peer review of final assessments to be used for management advice.

3. Approximately two weeks prior to the meeting documents containing the three stock assessments and other background material were made available and reviewed. Three CIE reviewers participated in the review during a meeting at the AFSC, Seattle from the 18^{th}-20^{th} April 2017. Presentations by AFSC staff were made covering the EBS , Slope and Aleutian Islands surveys, the observer program, the age determination program and the Amendment 80 fishery. Assessment authors also presented updated descriptions of the stock assessments. During the meeting, the Reviewer actively participated in the discussions and offered comments and suggestions to assist in improving the assessments.

Summary of findings

(i) General Comments

4. All three species assessments are based on a similar range of data types and modelling approaches. The principal data comprise biomass estimates obtained from the EBS shelf survey, the slope survey and the Aleutian Islands survey. The surveys provide a swept area biomass value with an associated sample design based CV. In addition, the surveys provide length and/or age compositions of the species concerned. During the review, presentations were made about the survey design and methodology. It is clear that

considerable effort has gone into standardization and the adoption of good practice through the development of quality control procedures. This includes ensuring tows are valid through monitoring net performance, adopting an appropriate survey design and optimizing sampling effort in response to fish distribution and abundance. In all cases, the surveys are carried out on commercial vessels and these change over time, which may give rise to changes in survey catchability and may be an important source of variability.

5. The three surveys all have somewhat different histories and this has resulted in the use of differing sampling gears and tow duration. Consequently, while the surveys provide notionally absolute estimates of biomass in the areas sampled, it is likely that there are in practice differences in the scale of the estimates making summing biomass across the shelf, slope and Aleutian Islands to obtain a global stock biomass value problematic. This difficulty affects all three assessments to some degree and is discussed under the individual species below.

6. An overview of the flatfish age reading program showed that good quality control procedures were in place to ensure consistent interpretation of growth rings. However, it is clear that age determination of older fish is challenging and there is a fairly high degree of disagreement between readers. It is likely, therefore, that age determination is much less certain at older ages. This may be relevant to the choice of plus group in the assessment especially when trying to fit the model to age compositions. A lower aged plus group might be more robust.

7. Typically, more otoliths are collected than are actually aged suggesting human resource limitations. Since age data are an information rich source, increasing the number of specimens read is likely to be a worthwhile investment. Although some otoliths have been collected from the arrowtooth and Kamchatka flounder fisheries, none have been read and the age information in these assessments is limited to the surveys.

8. Catch biomass data are important for the assessment as they effectively inform the model on the magnitude of fishing mortality rate. In the case of arrowtooth and Kamchatka flounders, these species were not distinguished in the catch records until 2011 and consequently the relative split between the species is subject to estimation error. From 1991-2006 the split is based on an assumption of 10% Kamchatka while for 2007-2010 some observer data has been used. The uncertainties associated with reconstructing the catches in this way is likely to have the greatest effect on the Kamchatka assessment as small errors in the proportions assumed can have a large effect on the estimated catches. As discussed below in the Kamchatka section below, there may be some advantage in doing a combined arrowtooth and Kamchatka assessment where the model estimates the catch split using the relative abundance of the species in the surveys.

9. Although there are differences in detail, the modelling approach adopted for the three assessments is similar. The population dynamics are described by an age structured model where annual recruitment is estimated as a random variable. Fishing mortality is formulated as the product of an age or size effect scaled by an "effort" or year effect, while natural mortality is usually an externally specified constant. In order to generate fitted length compositions to compare to observed values, the age structured population is converted to length using an age-length matrix based on a fixed growth curve and estimates of variance of length at age. Fixed growth is a strong assumption and may contribute to the problems in fitting the length compositions, but it is likely to be difficult to estimate variable growth within the model unless there is a clear signal in the length data.

10. The model parameters are estimated by fitting to the data using an objective function based on maximum likelihood, but with the ability to weight the data components according to user choice using "emphasis" factors. In addition, in each assessment the likelihood is the sum of components drawn from either a lognormal distribution (catch biomass, survey biomass) or a multinomial (length or age compositions) which makes correct weighting of the data a considerable challenge. Strictly speaking, each term in the likelihood should represent a true probability - the probability of the data given the parameters, θ; i.e. $p(\text{data}|\theta)$. However, for the multinomial the effective sample size is not really known making it unclear how much weight to give these data. Furthermore, the observed catch is typically treated as very precise even though, as in the case of Kamchatka flounder, these are not known with any precision in the earlier period.

11. In the case of the lognormal distributions the "emphasis" should be related to the observation error but the procedure is typically to use externally derived estimates of the sample error combined with an arbitrary weighting factor. For the surveys the observation error is regarded as the sample CV based on the survey design and this will almost certainly underestimate the true error since it omits many sources of variation such as changes of survey vessel, changes to fish availability and environmental factors such as weather. Depending on the information content in the data, in principle the observation error could be estimated within the model as the sum of the known sampling error, σ_{sample} and the unknown process error σ_{other}. Similarly if the observed catch is truly precise it should be possible to estimate its observation error within the model. In practice, if these variances cannot be estimated then it would suggest inconsistency between the differing sources of data. This is an issue which would merit further investigation to get a better understanding of uncertainty and to reduce exposure to arbitrary choice of weighting.

12. The assessment models all seek to estimate a relatively large number of parameters (between 85 and 166) that include annual recruitment deviations, annual fishing mortality deviations and selectivities for the surveys and the fishery. In some cases, survey catchability is also estimated as well as a temperature-catchability relationship. With this large number of parameters there is a danger of over-fitting the data and

arriving at a model with poor predictive power. There is a need for a systematic way to arrive at the most parsimonious model that provides adequate information to calculate management reference points and perform stock forecasts. Much of the discussion of the assessments revolved around the closeness of fit to the data and while this is an appropriate diagnostic, it does not address the question of parsimony. A conventional way forward might be to calculate an information statistic such as AIC, DIC or BIC for each model to investigate the trade-off between model complexity and goodness of fit. A particular area where this may be of use is whether it is worth trying to estimate separate selectivities by gender. Visual inspection of many of the gender specific selectivity plots suggested that these did not differ very much and that a single selectivity curve would suffice. Furthermore, if gender differences occur due to differential growth, then size alone may be adequate to explain gender differences in selectivity.

13. Although fisheries change over time, such change is often gradual both in terms of the overall level of fishing activity and the selectivity of the fleets. Fishing mortality or selectivity in a given year is therefore likely to be a good predictor of the same quantity a year ahead and this property could be exploited to reduce the effective number of parameters to be estimated in the model. It avoids the need to estimate separate selectivity parameters in blocks of years or to estimate a fishing mortality rate parameter for every year.

14. While each assessment document provides a description of the model and discusses the parameters estimated outside the model, it was still not clear which quantities were actually estimated and which were fixed or constrained externally. For example, in the Kamchatka flounder assessment were the σ^2_F and σ^2_R fixed or estimated? It would be desirable to provide a list of the model parameters to be estimated and their priors, and a list of constants set externally. The parameter list should also identify the variances that are estimated and those that are specified.

15. The current approach to model fitting lies somewhere between conventional maximum likelihood and Bayesian methods. Penalty functions on some parameters may be added to the data likelihood terms, such as the variance of recruitment deviations. The penalty functions are often referred to as “priors”, though their role is usually to constrain the location of maximum of the likelihood rather than obtain true posterior distributions of the estimated parameters. While such a distinction may appear somewhat philosophical, there is value in being more explicit about the choice of priors on all parameters and examining their posterior distributions in order to understand the true contribution of information from the data versus the analyst’s perception of prior knowledge. Even where priors are assumed uniform, making this explicit with stated bounds, where applied, would greatly assist in transparency as it would clearly identify model parameters and the assumptions on any prior knowledge.

16. The need to examine posterior distributions more systematically is illustrated in the flathead sole assessment (Figure 9.12 in McGilliard et al., 2016) where MCMC has been used to sample the posteriors of some parameters. These show that some posteriors are multimodal which makes interpretation of a maximum likelihood value problematic. Pure maximum likelihood has the advantage of computational speed but multimodality in the posteriors illustrates the weakness of relying on asymptotic estimates of parameter variance and the mode as the best measure of location. Current gradient based MCMC samplers can be very efficient (Hoffman and Gelman, 2014) and it may thus be preferable to sample from the posteriors of all the parameters from the outset rather than rely on maximizing the likelihood to estimate the parameters and their asymptotic variances.

17. Adopting a more formal Bayesian approach would overcome many of the problems discussed above as it would:

a. Clearly identify the parameters to be estimated and their prior distributions
b. Identify constants in the model
c. Enable the analyst to assess the information in the data by comparing priors to posteriors
d. Test the reliability of the model fit using conventional MCMC chain convergence statistics
e. Provide more realistic estimates of parameter uncertainty
f. Facilitate the estimation of uncertainty in derived quantities such as SSB and reference points.

18. A commonly used model diagnostic is retrospective analysis where data at the end of the time series are progressively removed and the model refit to compare estimated trends in biomass and fishing mortality. This has been used in the assessments reviewed and is an important and useful tool in testing model consistency. It is worth remembering that ultimately the assessment model will be used to calculate reference points and make a forward projection for the purposes of ABCs. Given that most stock assessment models are over-parameterised (usually by necessity) they may fit the data well, but have poor ability to predict into the future. In performing the retrospective analysis, it is also important to calculate reference points and make a forward projection (perhaps two years ahead) to test the consistency of the model in reference point calculation and its ability to forecast.

(ii) Bering Sea and Aleutian Islands Arrowtooth flounder

Evaluation of the ability of the stock assessment model for arrowtooth flounder, combined with the available data, to provide parameter estimates to assess the current status of arrowtooth flounder in the Bering Sea and Aleutian Islands.

19. The assessment model is parameter rich whereas the data are of moderate amount. Perhaps the main limitation from the assessment model perspective is the absence of age composition data for the fishery which will mean that the fishery selectivity estimate is more uncertain since it relies on the length compositions alone. Runs performed during the review meeting showed that the fishery selectivity was sensitive to the relative weights

given to the three surveys. Uncertainty in the fishery selectivity will have a direct bearing on the calculation of reference points.

20. The shelf survey (EBS) is believed to account for about 80% of the total stock biomass and assuming the swept area estimates of biomass are reasonably realistic this results in very low estimates of exploitation rate (4% or less). If correct, then the stock dynamics will be almost entirely driven by recruitment and natural mortality. The EBS biomass estimates show an overall increase in stock biomass over the assessment period, though it has fluctuated without long-term trend since 1995. If fishing mortality is as low as estimated, then the current biomass will be very close to $B_{100\%}$ and this is what the preliminary assessment shows.

21. For the assessment to alter the perception of stock status it would require substantially lower biomass estimates from the survey and hence the reliability of the swept area calculation is crucial. Currently, the uncertainty in the biomass estimates is quantified only in the sample variance which will significantly underestimate the true uncertainty. An important area for further analysis is to investigate the contribution of scaling values used to derive the swept area to the overall uncertainty. It is likely that the probability of a fish in the path of the survey gear being retained will have a major influence on the biomass estimates and needs to be evaluated.

Evaluation of the strengths and weaknesses in the stock assessment model for arrowtooth flounder.

22. The main strength of the model is that it is able to make use of both biomass, age and length composition data in a unified framework. There is also a fairly well established statistical framework in which to estimate the parameters. The model is supported by comprehensive survey biomass estimates which should provide high quality estimates of biomass trends.

23. Because the model is data rich but there is relatively little contrast in the data, some of the parameters will not be well estimated, and this is apparent in the sensitivity of the selectivity estimates to changes in the relative weight given to the surveys. It would be desirable to try to reduce the number of parameters to avoid over-fitting the data.

24. The preliminary assessment included a data component incorporating sex ratio data based on the survey. There are two problems with this. Firstly, it is in effect using the data twice since the survey data used in the model are already separated by sex, and secondly a lognormal distribution was assumed which is not appropriate for proportion data. If sex ratio data are used, then they should be transformed to logits before using a normal distribution.

25. In fitting the model, both natural mortality and survey catchability are fixed. These quantities both affect the scale of the biomass estimates and need to be consistent to avoid bias emerging in the estimates of other parameters. This is likely to be in the selectivity

parameters as these reconcile the modelled biomass with the observed values. It might be worth investigating a model run with M fixed but survey catchability estimated freely.

Evaluation of the assumption that male natural mortality is higher than female in arrowtooth flounder.

26. Smaller individuals in a population are generally considered to suffer higher natural mortality due to their greater vulnerability to predation and other factors affecting survival. Arrowtooth males clearly grow more slowly than females so would be expected to have, on average, a higher M than similarly aged females. In the assessment runs, likelihood profiling was done over male M when female M was fixed at a range of values. In all cases the male Ms were higher than females' and this is consistent with theory. From a modelling perspective, the issue is whether the difference in M is due to gender or actually to body size. If the latter, then modelling M by size might be simpler and would account for differential survival by size rather than assuming a fixed value for all ages. The relationship described by Lorenzen (1996) is widely used in stock assessments and may be relevant here.

Recommendations for further improvements to the assessment model.

27. During the review meeting, there was discussion on the extent to which there were distinct selectivities for males and females. Visual inspection of selectivity curves, and sensitivity runs changing the relative catchability for each survey suggest that in practice it may not be necessary to estimate separate selectivites for males and females. Size alone may be an adequate co-variate. There is, in general, a need to try to find the most parsimonious model given the limitations of the data.

28. Prior to 2011 catch records for arrowtooth also included Kamchatka flounder. This is probably of little significance for arrowtooth assessment, but is relevant to the Kamchatka flounder and there may be benefits to be gained by doing a combined assessment as described in the Kamchatka flounder section.

29. As presently structured, the assessment area includes both the Bering sea and the Aleutian Islands. One problem with choosing the combined area is that the AI survey differs from the EBS and Slope survey and requires an assumption about the relative contribution of the AI survey to the total biomass. Getting this wrong will result in an inappropriate weight being given to all three surveys and biased selectivity estimates. In practice, if the Aleutian Island area could be regarded as a separate assessment unit, it may simplify some of the necessary assumptions and might be an avenue worth exploring. Clearly a potential downside is the lower frequency of the AI survey which might undermine a standalone assessment.

30. It was not clear how the temperature-survey catchability relationship was handled in the assessment. Figure 6.8 in Spies et al. (2016) shows a very close relationship but I assume this is simply q calculated from deterministic equation using bottom temperature. If such a relationship exists, it should be estimated within the model. However, unless the effect is very strong, there is a danger of over-fitting the data and the need for these additional parameters needs rigorous evaluation.

31. As mentioned earlier the sex ratio data should be removed from the likelihood.

(iii) Bering Sea and Aleutian Islands Kamchatka flounder

Evaluate stock assessment approach to model the Kamchatka flounder resource using three spatially distinct trawl surveys to provide reliable estimates of productivity, stock status, and statistical uncertainty for management advice.

32. As with the arrowtooth flounder, the assessment model is parameter rich whereas the data are of moderate amount, perhaps even poor, given the issues with mixed catches of arrowtooth. Perhaps the main limitation from the assessment model perspective is the absence of age composition data for the fishery which will mean that the fishery selectivity estimate is more uncertain since it relies on the length compositions alone. Uncertainty in the fishery selectivity will have a direct bearing on the calculation of reference points. It required the slope to be fixed to obtain a credible model fit which is indicative for poor information in the data.

33. There were clearly problems in fitting the model to obtain good estimates of survey catchability, M and selectivity simultaneously. Trying to estimate M and q is fraught with difficulty since both affect the scale of the biomass and there is a need for the data to contain information on scale. This could be provided by the catch, but since catch appears to be a very small proportion of the biomass, such information is lacking. Furthermore, the catch data prior to 2011 are to varying degrees contaminated by assumptions about the arrowtooth catch.

34. Examination of the estimated stock biomass shows a more or less continuous upward trend over the assessment period but this is not fully reflected in any of the individual the survey biomass trends and there are clearly systematic residuals in the fitted survey values. The 95% CI for the biomass suggest that it is just above $B_{40\%}$ with the lower bound on the reference point. The confidence interval is presumably based on the asymptotic variance estimates and appears far too narrow given the uncertainties in the data, the need for strong constraints on fishery selectivity and catchability, and the poor quality of the catch data. It would seem that stock status is uncertain.

35. A possible alternative approach to the assessment would be to recognize that this species is a component of the fishery for arrowtooth and develop an assessment model that combined both species. This could be done by retaining separate species identity in the data that are of known species, but fit to combined catches where the species are not distinguished. This would avoid the need to apply an *ad hoc* Kamchatka catch proportion to the early catch series. The approach would be to model fully selected F as a time series common to both stocks but with a random effect added to account for a species specific fishing mortality on Kamchatka flounder. For example, if F_{AT} is the fully selected F on arrowtooth and F_K is the fishing mortality on Kamchatka, one might consider the following model:

$$F_{AT,t} = F_{AT,t-1} \exp(\varepsilon_t), \qquad \varepsilon_t \sim Normal(0, \sigma_{AT})$$

$$F_{K,t} = F_{AT,t} \exp(\epsilon_t + \alpha\epsilon_{t-1}), \quad \epsilon_t \sim Normal(0, \sigma_K)$$

36. These equations can be used to generate species specific catches that would enter the likelihood either as a combined catch or individual species catches. If necessary, a catch proportion parameter could be estimated and also propagated as a time series.

Evaluate likelihood profile approach to estimate natural mortality rate (and suggest/provide alternatives?)

37. As is mentioned in the assessment document (Wilderbuer et al., 2016) M and q tend to be confounded. Some practitioners believe they know more about M so fix this value and estimate q. Others take an alternative view and fix q so that M can be estimated. It is very much a choice depending on what you believe you know about these two quantities. The many studies that provide a basis for choosing M use meta-analyses or life history theory so one might argue that there are evidential grounds to identify an appropriate value of M without trying to estimate it from the data. Furthermore, the best chance of estimating M from within the assessment occurs when there is strong contrast in the data. This does not appear to be the case with Kamchatka flounder, especially as F appears to be very low.

38. Clearly much work has been done in the region to derive estimates of total biomass from surveys and implicitly this means q=1. But the attempts to estimate M from the assessment model appear very uncertain. In these circumstances, it is probably worth fixing M externally. However, if F really is as low as the assessment suggests, total mortality, Z, will be a good indicator of the magnitude of M. Hence, plotting log numbers from the survey against age for each cohort and calculating the slope might provide an adequate measure of the magnitude of M.

39. If there is a desire to try to estimate M from the model, one alternative approach is to assume it is size dependent and use the Lorenzen (1996) equation as the basis for a prior. For example, Cook et al. (2015) used priors on the two parameter Lorenzen equation to estimate M for Atlantic cod. The approach requires a full Bayesian analysis and has been discussed earlier in the General Comments section. A size dependent M might overcome the need to have separate values by sex as discussed in the arrowtooth assessment.

Evaluate how survey catchability estimates are derived based on assumptions about relative stock distributions.

40. Clearly this is a difficult problem. While it is to some degree evading the issue, separating out the Aleutian Islands as a single assessment unit might simplify the assumptions that have to be made. This is referred to in the arrowtooth review above. For the remaining area some analysis of the linkage between the shelf and slope populations may be fruitful. There appears to be evidence to suppose that smaller fish predominantly on the shelf move to the

slope as they get larger and older. Understanding this movement may offer a means of calibrating the surveys so that a single biomass estimate can be obtained.

(iv) Bering Sea and Aleutian Islands flathead sole

Evaluation of the ability of the stock assessment model for flathead sole, with the available data, to provide parameter estimates to assess the current status of flathead sole in the Bering Sea and Aleutian Islands

41. The flathead sole assessment model is similar in most respects to the models used for Kamchatka sole and arrowtooth flounder. It is parameter rich and uses the same likelihood approach. Hence many of the general comments made to the previous assessments relating to the statistical approach apply here too. This assessment does include age data for the fishery which should improve the ability of the model to estimate selectivity for the fishery.

42. Survey data used include the AI survey and the EBS shelf survey. In the model, however, the AI and EBS survey are combined into a single survey series using a regression approach. The AI survey represents a very small component of the total biomass so the abundance data will be dominated by the EBS survey.

43. Model runs presented in the assessment document show a sharp decline in fishing mortality in the late 1970s associated with a strong increase in stock biomass. Subsequently, the fishing mortality stabilizes at a low level and stock biomass reaches a peak in the early 1990s followed by a gradual decline. Superficially, the precipitous decline in fishing mortality looks unrealistic and may well be an artifact of the model as discussed below.

44. The assessment document provided before the review meeting discussed the currently accepted model updated with more recent data. Concerns have been expressed about the current model which relate to:
    - Modelling selectivity curves
    - Lack of fit to fishery length and age compositions
    - Modelling a temperature-catchability relationship is questionable

45. These concerns and the fact that the model code is considered inflexible has resulted in work to transfer the current model into Stock Synthesis (SS) before further development. During the review meeting a presentation by the lead analyst summarized the current status of that exercise. Much of the work involved demonstrating that the current model can be accurately reproduced in SS. At the time of the meeting, it appeared that the SS version was able to reproduce most of the original model stock biomass and recruitment estimates, but that further work was still required to fully reconcile fishing mortality estimates.

46. In view of the current state of model development, it is perhaps premature to evaluate the ability of the model to assess the status of the stock.

Evaluation of the strengths and weaknesses in the stock assessment model for Bering Sea/Aleutian Islands (BSAI) flathead sole

47. The main strength of the model is that it is able to make use of both biomass, age and length composition data in a unified framework. There is also a fairly well established statistical framework in which to estimate the parameters. The model is supported by comprehensive survey biomass estimates which will provide high quality estimates of biomass trends. The additional age composition data for the fishery will also strengthen this assessment.

48. There may be some value in reconsidering the use of survey abundance time series. In practice, the current assessment effectively only uses the EBS survey as the AI component is so small. As discussed for the other flatfish, there may be a case for treating the AI as a separate assessment unit where the AI survey could be used as the principal fishery independent data. As things currently stand any signal in the AI survey will be obliterated by the EBS survey. There is also data available from the slope survey which might provide better information on larger and older fish.

Evaluation of alternatives to the current length-based survey selectivity curves used in the assessment

49. An important difference between this assessment and the arrowtooth/Kamchatka flounder assessments is that age composition data are available for both the surveys and the fishery from the mid-1990s. This provides the assessment with much more information than length data alone. While it sounds somewhat heretical to say so, I do wonder if the length frequency data really provide much useful information and whether omitting size data would materially degrade the assessment. Differences in growth rate and the problems of characterizing variance of length at age conspire to make fitting length data extremely uncertain and this is already apparent in current model runs. The EBS survey has 12 years of age composition data spanning the period 1982-2015 which should be adequate to estimate age specific selectivity reasonably well. If so then it should be possible to configure the model to be based on age selectivity alone. For years when no age composition data are available, the survey numbers (without size differentiation) could be used as data to inform the model about total abundance. Models that do not distinguish size but operate only on abundance can perform well (Mesnil, 2003).

Potential evaluation of an equivalent BSAI flathead sole assessment model in Stock Synthesis

50. During the review meeting a SS version of the model was presented and discussed. In the time available and without a fully documented assessment, it is difficult to offer a considered evaluation of the current SS implementation. Clearly, the SS framework offers the potential to investigate a range of alternative models and selectivity is an area that merits further work. Runs with fishery selectivity fixed, divided into blocks or allowed to vary with time were presented. Increasing the flexibility of the model to fit the data by allowing more freedom in selectivity runs the risk of over-fitting the model. There are good reasons to allow at least two time blocks due to management changes in the fishery but allowing annual deviations in selectivity without some constraint is probably unwise. As discussed in the general comments, modelling the selectivity parameters as an

autocorrelated time series may help in adding some stiffness to the model and reduce the number of effective parameters to be estimated.

51. Two notable features emerge from the various selectivity assumptions. Firstly, the strong decline in fishing mortality in the ADMB model largely disappears in the SS model. Secondly, as can be seen in slide 66 of the BSAI Flathead Sole Complex presentation, estimates of fishing mortality prior to 1990 are highly sensitive to the choice of selectivity assumption. This is indicative that the early period is not well determined while more recent values are more robustly estimated.

52. It is proposed also to use the SS framework to estimate both q and M. One should be realistic about the limitations of the data and think carefully about what can be estimated. Both these quantities are scaling factors on the biomass and are unlikely to be uniquely identifiable without constraints. It is worth remembering that M is size dependent in the real world and while a single global value may be estimable from the model, it may simply be the dumping ground for a range of other hidden model miss-specification.

(v) NMFS review process

53. The review was conducted in a constructive atmosphere. Background information presented at the meeting on the input data was, comprehensive, extremely useful, and helped in evaluating the assessments. Relevant documents were made available two weeks prior to the meeting which provided important preparatory material. However, by the time of the meeting, further work on the assessments had been done which meant that consideration of revisions could only be fairly superficial in the time available as the only documentation was in PowerPoint presentations.

Conclusions and Recommendations

54. The three assessments reviewed are all at an intermediate stage of development and there are a number of possible options for refinement before final assessments are made. All are based on a similar data and modelling approach which is well established and reflects the Stock Synthesis school of assessment. The split by sex in the assessments potentially increases the number of parameters to be estimated if sex-specific selectivity curves are modelled. There is a danger of over-fitting the data. I would recommend that efforts are made to reduce the number of parameters by consolidating selectivity curves were possible. Modelling fully selected F as a time series may also help to reduce the effective number of parameters. An information statistic such as AIC or DIC may help guide model selection.

55. I felt that there was lack of clarity about the model quantities that were treated as known or as parameters to be estimated. This also applies to the use of informative priors or emphasis factors. To some degree this is the result of a somewhat informal likelihood approach without a clear statistical rationale for priors or penalty function distributions.

Consequently, it is hard to evaluate if the data in the likelihood are receiving the correct weighting and therefore how to interpret the estimated variances on the quantities of interest. The modelling environment used has, in my view, reached a stage where statistical rigor could be improved by formally adopting a full Bayesian approach where priors are chosen to reflect accurately was is known before the data are used. Where highly informative priors are used (such as on the catch data even when this is uncertain) in order to force a maximum likelihood solution, this needs to be made explicit in order to understand the true uncertainty in the assessment.

56. The three research vessel surveys are perhaps the most important data source in the assessments as they provide data split by species and sex as well as providing age composition information sometimes absent for the fishery. Since the fishery appears to take a small proportion of the biomass, the surveys will provide the most information about stock biomass trends. All three surveys appear to be conducted to a high standard but their Achilles heel is their geographical separation and use of differing sampling protocols which makes combining the biomass estimates difficult. Arguably there is much to be gained by being able to combine estimates from the EBS and slope surveys since this is a contiguous area and species are distributed in both. A study to find ways of inter-calibrating the survey biomass estimates might help in avoiding the need to make ad hoc fixes in the assessment.

57. The AI survey takes place in a somewhat separate area and there may be a case for performing separate assessments for this area to avoid the assessment being dominated by the EBS survey that may not appropriately characterize the populations in the Aleutian Islands.

58. As the surveys take place on commercial vessels that change periodically, an investigation to vessel effects would help in understanding if changes in catchability are of material importance. The assessment approach is to assume survey catchability is constant and in some cases known which makes understanding the effect of vessel changes particularly important. In addition, the current estimates of survey biomass variance are based only on the sample design and will underestimate the overall uncertainty. A study to obtain a more comprehensive estimate of survey precision would be desirable as this directly affects the relative weight given to the data in the likelihood of the assessment model.

References

Cook, R., Holmes, S. and Fryer, R. 2015. Grey seal predation impairs recovery of an over-exploited fish stock. Journal of Applied Ecology 52, 969–979. doi: 10.1111/1365-2664.12439

Hoffman, M. D. and Gelman, A. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15, 1351-1381.

Lorenzen, K., 1996. The relationship between body weight and natural mortality in juvenile and adult fish: a comparison of natural ecosystems and aquaculture. Journal of Fish Biology, 49: 627–647.

Mesnil, B. (2003). The Catch-Survey Analysis (CSA) method of fish stock assessment: an evaluation using simulated data. Fisheries Research 63, 193–212.

Appendix 1: Bibliography of materials provided for review

Documents provided before the review meeting

Spies, I., Wilderbuer, T.K., Nichol, D.G. and Hoff, J., Palsson, W., 2016. Arrowtooth flounder. *Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions*, pp.921-1012. http://www.afsc.noaa.gov/REFM/Docs/2016/BSAIatf.pdf

Doyle, M., Debenham, C., Barbeaux, S., Buckley, T., Spies, I., Pritle, J., Shotwell, K., Wilston, M., Cooper, D., Stockhausen, W., and Duffy-Anderson, J. In Prep. A full life history synthesis of Arrowtooth Flounder ecology in the Gulf of Alaska.

Wilderbuer, T. and Turnock, B. 2009. Sex-Specific Natural Mortality of Arrowtooth Flounder in Alaska: Implications of a Skewed Sex Ratio on Exploitation and Management, North American Journal of Fisheries Management, 29:2, 306-322, DOI: 10.1577/M07-152.1.

Wilderbuer, T., J. Ianelli, D. Nichol, and R. Lauth. 2016. Assessment of the Kamchatka flounder stock in the Bering Sea and Aleutian Islands. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions. North Pacific Fisheries Management Council, Anchorage, AK. http://www.afsc.noaa.gov/REFM/Docs/2016/BSAIkamchatka.pdf

NPFMC. 2017. BSAI Introduction. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions. North Pacific Fisheries Management Council, Anchorage, AK. http://www.afsc.noaa.gov/REFM/Docs/2016/BSAIintro.pdf

McGilliard, C.R., Nichol, D. and Palsson, W. 2016. 9. Assessment of the Flathead Sole-Bering flounder Stock in the Bering Sea/Aleutian Islands Regions. In Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea and Aleutian Islands. pp. 1229-1318. North Pacific Fishery Management Council, P.O. Box 103136, Anchorage, AK 99510.

Baker,M.R.,Hollowed,A.B.,Delineating ecological regions in marine systems: Integrating physical structure and community composition to inform spatial management in the eastern Bering Sea. Deep Sea Res.II(2014),ttp://dx.doi.org/10.1016/j.dsr2.2014.03.001i

Powerpoint presentations

Delsa Anderl & Beth Matta Flatfish Age Determination at the Alaska Fisheries Science Center

Ingrid Spies, Thomas K. Wilderbuer, Daniel G. Nichol, Jerry Hoff, Wayne Palsson, James N. Ianelli, Andy Kingham, Ren Narita . The Bering Sea and Aleutian Islands, and Gulf of Alaska arrowtooth flounder stock assessment.

Bob Lauth. **Eastern Bering Sea shelf bottom trawl survey of groundfish and crab resources**

Alan Haynie. The Amendment 80 fishery: A long-term view on management changes that have impacted this North Pacific multi-species fishery.

Gerald Hoff. Eastern Bering Sea Upper Continental Slope Groundfish Bottom Trawl Survey.

Thomas Wilderbuer, James Ianelli, Daniel Nichol and Robert Lauth. **Assessment of the Kamchatka Flounder stock in the Bering Sea and Aleutian Islands**

Ned Laman, Christina Conrath, Sean Rooney, Peter Munro, Jay Orr, Nate Raring, Chris Rooper, Paul von Szalay, Mark Zimmermann, Susanne McDermott, Wayne Palsson. Aleutian Islands Bottom Trawl Survey.
(1980 – present)

Bob Lauth, Lyle Britt, Jason Conner, Rebecca Haehn, Jerry Hoff, Elaina Jorgensen, Stan Kotwicki, Dan Nichol, Elizabeth Perkins, Duane Stevenson, & Cynthia Yeung. Eastern Bering Sea Shelf Bottom Trawl Survey: Flatfish CIE Review.

Carey McGilliard, Dan Nichol, and Wayne Palsson. BSAI Flathead Sole Complex.

Tom Wilderbuer. Overview of Bering Sea flatfish fisheries and management.

Appendix 2: Statement of Work

Statement of Work

National Oceanic and Atmospheric Administration (NOAA)
National Marine Fisheries Service (NMFS)
Center for Independent Experts (CIE) Program
External Independent Peer Review

Fisheries Stock Assessments for Arrowtooth Flounder, Flathead Sole and Kamchatka Flounder

Background

The National Marine Fisheries Service (NMFS) is mandated by the Magnuson-Stevens Fishery Conservation and Management Act, Endangered Species Act, and Marine Mammal Protection Act to conserve, protect, and manage our nation's marine living resources based upon the best scientific information available. NMFS science products, including scientific advice, are often controversial and may require timely scientific peer reviews that are strictly independent of all outside influences. A formal external process for independent expert reviews of the agency's scientific products and programs ensures their credibility. Therefore, external scientific peer reviews have been and continue to be essential to strengthening scientific quality assurance for fishery conservation and management actions.

Scientific peer review is defined as the organized review process where one or more qualified experts review scientific information to ensure quality and credibility. These expert(s) must conduct their peer review impartially, objectively, and without conflicts of interest. Each reviewer must also be independent from the development of the science, without influence from any position that the agency or constituent groups may have. Furthermore, the Office of Management and Budget (OMB), authorized by the Information Quality Act, requires all federal agencies to conduct peer reviews of highly influential and controversial science before dissemination, and that peer reviewers must be deemed qualified based on the OMB Peer Review Bulletin standards.
(http://www.cio.noaa.gov/services_programs/pdfs/OMB_Peer_Review_Bulletin_m05-03.pdf).
Further information on the CIE program may be obtained from www.ciereviews.org.

Scope

The Alaska Fisheries Science Center's (AFSC) Resource Ecology and Fisheries Management Division (REFM) requests an independent review of the integrated stock assessments that have been developed for three Bering Sea flatfish species; arrowtooth flounder, flathead sole and Kamchatka flounder. The fishery for these species is managed by the North Pacific Fisheries Management Council. The sum of the Allowable Biological Catches (ABCs) for these three species is 142,529 t in 2017, with catch levels annually set lower than the ABC due to a 2.0 million t harvest cap for all species and constraints due to Pacific halibut bycatch limits and markets. The catch limits are established using Automatic Differentiation (AD) Model software that uses survey abundance data and survey and fishery age and length composition data with a harvest control rule to model the status and productivity of these stocks and set quotas. Having these assessments vetted by an independent expert review panel is a valuable part of the AFSC's review process. The Terms of Reference (TORs) of the peer review and the tentative agenda of the meeting are below.

Requirements for CIE Reviewers

NMFS requires three CIE reviewers to conduct an impartial and independent peer review in accordance with the SOW, OMB Guidelines, and the TORs below. The reviewers shall have working knowledge and recent experience in the application of fisheries stock assessment processes and results, including population dynamics, separable age-structured models, harvest strategies, survey methodology, and the AD Model Builder programming language. Experience with the Stock Synthesis Assessment Model would also be helpful. They should also have experience conducting stock assessments for fisheries management.

Statement of Tasks

- Review the following background materials and reports prior to the review meeting:

Spies, I., Wilderbuer, T.K., Nichol, D.G. and Hoff, J,. Palsson, W., 2016. Arrowtooth flounder. *Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions*, pp.921-1012.
http://www.afsc.noaa.gov/REFM/Docs/2016/BSAIatf.pdf

Doyle, M., Debenham, C., Barbeaux, S., Buckley, T., Spies, I., Pritle, J., Shotwell, K., Wilston, M., Cooper, D., Stockhausen, W., and Duffy-Anderson, J. In Prep. A full life history synthesis of Arrowtooth Flounder ecology in the Gulf of Alaska.

Wilderbuer, T. and Turnock, B. 2009. Sex-Specific Natural Mortality of Arrowtooth Flounder in Alaska: Implications of a Skewed Sex Ratio on Exploitation and Management, North American Journal of Fisheries Management, 29:2, 306-322, DOI: 10.1577/M07-152.1.

Wilderbuer, T., J. Ianelli, D. Nichol, and R. Lauth. 2016. Assessment of the Kamchatka flounder stock in the Bering Sea and Aleutian Islands. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions. North Pacific Fisheries Management Council, Anchorage, AK.
http://www.afsc.noaa.gov/REFM/Docs/2016/BSAIkamchatka.pdf

NPFMC. 2017. BSAI Introduction. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions. North Pacific Fisheries Management Council, Anchorage, AK.
http://www.afsc.noaa.gov/REFM/Docs/2016/BSAIintro.pdf

McGilliard, C.R., Nichol, D. and Palsson, W. 2016. 9. Assessment of the Flathead Sole-Bering flounder Stock in the Bering Sea/Aleutian Islands Regions. In Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea and Aleutian Islands. pp.

1229-1318. North Pacific Fishery Management Council, P.O. Box 103136, Anchorage, AK 99510. http://www.afsc.noaa.gov/REFM/Docs/2016/BSAIflathead.pdf

- Attend and participate in the panel review meeting
    - The meeting will consist of presentations by NOAA and other scientists, stock assessment authors and others to facilitate the review, to provide any additional information required by the reviewers, and to answer any questions from reviewers
- After the review meeting, reviewers shall conduct an independent peer review in accordance with the requirements specified in this SOW, OMB guidelines, and TORs, in adherence with the required formatting and content guidelines; reviewers are not required to reach a consensus
- Each reviewer may assist the Chair of the meeting with contributions to the summary report, if required by the TORs
- Deliver their reports to the Government according to the specified milestone dates

Foreign National Security Clearance

When reviewers participate during a panel review meeting at a government facility, the NMFS Project Contact is responsible for obtaining the Foreign National Security Clearance approval for reviewers who are non-US citizens. For this reason, the reviewers shall provide requested information (e.g., first and last name, contact information, gender, birth date, passport number, country of passport, travel dates, country of citizenship, country of current residence, and home country) to the NMFS Project Contact for the purpose of their security clearance, and this information shall be submitted at least 30 days before the peer review in accordance with the NOAA Deemed Export Technology Control Program NAO 207-12 regulations available at the Deemed Exports NAO website: http://deemedexports.noaa.gov/ and http://deemedexports.noaa.gov/compliance_access_control_procedures/noaa-foreign-national-registration-system.html. The contractor is required to use all appropriate methods to safeguard Personally Identifiable Information (PII).

Place of Performance

The place of performance shall be at the contractor's facilities, and at the Alaska Fisheries Science Center, Seattle, Washington.

Period of Performance

The period of performance shall be from the time of award through June 12, 2017. Each reviewer's duties shall not exceed 14 days to complete all required tasks.

Schedule of Milestones and Deliverables: The contractor shall complete the tasks and deliverables in accordance with the following schedule.

Within two weeks of award	Contractor selects and confirms reviewers
No later than April 4, 2017	Contractor provides the pre-review documents to the reviewers
April 18-20, 2017	Panel review meeting
May 8, 2017	Contractor receives draft reports
May 30, 2017	Contractor submits final reports to the Government

Applicable Performance Standards
The acceptance of the contract deliverables shall be based on three performance standards: (1) The reports shall be completed in accordance with the required formatting and content (2) The reports shall address each TOR as specified (3) The reports shall be delivered as specified in the schedule of milestones and deliverables.

Travel
All travel expenses shall be reimbursable in accordance with Federal Travel Regulations (http://www.gsa.gov/portal/content/104790). International travel is authorized for this contract. Travel is not to exceed $10,000.

Restricted or Limited Use of Data
The contractors may be required to sign and adhere to a non-disclosure agreement.

NMFS Project Contact:
Tom Wilderbuer
Tom.Wilderbuer@noaa.gov
National Marine Fisheries Service,
7600 Sand Point Way, NE, Bldg. 4,
Seattle, WA 98115-6349
Phone: (206) 526-4224

Peer Review Report Requirements

1. The report must be prefaced with an Executive Summary providing a concise summary of the findings and recommendations, and specify whether or not the science reviewed is the best scientific information available.

2. The report must contain a background section, description of the individual reviewers' roles in the review activities, summary of findings for each TOR in which the weaknesses and strengths are described, and conclusions and recommendations in accordance with the TORs.

   a. Reviewers must describe in their own words the review activities completed during the panel review meeting, including a brief summary of findings, of the science, conclusions, and recommendations.

   b. Reviewers should discuss their independent views on each TOR even if these were consistent with those of other panelists, but especially where there were divergent views.

   c. Reviewers should elaborate on any points raised in the summary report that they believe might require further clarification.

   d. Reviewers shall provide a critique of the NMFS review process, including suggestions for improvements of both process and products.

   e. The report shall be a stand-alone document for others to understand the weaknesses and strengths of the science reviewed, regardless of whether or not they read the summary report. The report shall represent the peer review of each TOR, and shall not simply repeat the contents of the summary report.

3. The report shall include the following appendices:

   Appendix 1: Bibliography of materials provided for review
   Appendix 2: A copy of this Statement of Work
   Appendix 3: Panel membership or other pertinent information from the panel review meeting.

Terms of Reference for the Peer Review

Bering Sea and Aleutian Islands Arrowtooth flounder

59. Evaluation of the ability of the stock assessment model for arrowtooth flounder, combined with the available data, to provide parameter estimates to assess the current status of arrowtooth flounder in the Bering Sea and Aleutian Islands.
60. Evaluation of the strengths and weaknesses in the stock assessment model for arrowtooth flounder.
61. Evaluation of the assumption that male natural mortality is higher than female in arrowtooth flounder.
62. Recommendations for further improvements to the assessment model.

Bering Sea and Aleutian Islands Kamchatka flounder

1. Evaluate stock assessment approach to model the Kamchatka flounder resource using three spatially distinct trawl surveys to provide reliable estimates of productivity, stock status, and statistical uncertainty for management advice.
2. Evaluate likelihood profile approach to estimate natural mortality rate (and suggest/provide alternatives?)
3. Evaluate how survey catchability estimates are derived based on assumptions about relative stock distributions.

Bering Sea and Aleutian Islands flathead sole

1. Evaluation of the ability of the stock assessment model for flathead sole, with the available data, to provide parameter estimates to assess the current status of flathead sole in the Bering Sea and Aleutian Islands
2. Evaluation of the strengths and weaknesses in the stock assessment model for Bering Sea/Aleutian Islands (BSAI) flathead sole
3. Evaluation of alternatives to the current length-based survey selectivity curves used in the assessment
4. Potential evaluation of an equivalent BSAI flathead sole assessment model in Stock Synthesis

Tentative Agenda

TBD

Alaska Fisheries Science Center

7600 Sand Point Way NE

Seattle, WA 98115

April 18-20, 2017 9AM - 5PM

Point of contact: Tom Wilderbuer (tom.wilderbuer@noaa.gov)

Appendix 3: Panel membership

CIE Reviewers

Robin Cook, University of Strathclyde
Sven Kupschus, CEFAS
Kevin Stokes, Consultant

Participants

Anne Hollowed (Chair, pt)	AFSC Status of stocks
Carey McGilliard	AFSC Status of stocks
Ingrid Spies	AFSC Status of stocks
Meaghan Bryan	AFSC Status of stocks
Tom Wilderbuer	AFSC Status of stocks
Sandra Lowe (Chair, pt)	AFSC Status of stocks
Jim Ianelli	AFSC Status of stocks
Alan Haynie	AFSC Economics program
Jerry Hoff	AFSC Bering Sea groundfish survey
Bob Lauth	AFSC Bering Sea groundfish survey
Dan Nichol	AFSC Bering Sea survey program
Ned Laman	AFSC Aleutian Islands groundfish survey
Beth Matta	AFSC Age and growth program
Delsa Anderl	AFSC Age and growth
Marlon Concepcion	AFSC Observer program
Todd Loomis	Industry (Ocean Peace)